Standardized Approach for Developing Probabilistic Exposure Factor Distributions

*Randy L. Maddalena[1], Thomas E. McKone[1,2] and Michael D. Sohn[1]

(1) Indoor Environment Department , Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California, 94720, USA

(2) School of Public Health, University of California, Berkeley, California, 942720

Corresponding Author: *Randy L. Maddalena
Indoor Environment Department
Lawrence Berkeley National Laboratory
1 Cyclotron Road (mail stop 90-3058)
Berkeley, California, 94720, USA
rlmaddalena@lbl.gov
Phone: 209) 962-5680
Fax: 510) 486-6658

March 2003

Report number: LBNL-52203

Journal: *Risk Analysis*

ABSTRACT

The effectiveness of a probabilistic risk assessment (PRA) depends critically on the quality of input information that is available to the risk assessor and specifically on the probabilistic exposure factor distributions that are developed and used in the exposure and risk models. Deriving probabilistic distributions for model inputs can be time consuming and subjective. The absence of a standard approach for developing these distributions can result in PRAs that are inconsistent and difficult to review by regulatory agencies. We present an approach that reduces subjectivity in the distribution development process without limiting the flexibility needed to prepare relevant PRAs. The approach requires two steps. First, we analyze data pooled at a population scale to (i) identify the most robust demographic variables within the population for a given exposure factor, (ii) partition the population data into subsets based on these variables, and (iii) construct archetypal distributions for each subpopulation. Second, we sample from these archetypal distributions according to site- or scenario-specific conditions to simulate exposure factor values and use these values to construct the scenario-specific input distribution. It is envisaged that the archetypal distributions from step 1 will be generally applicable so risk assessors will not have to repeatedly collect and analyze raw data for each new assessment. We demonstrate the approach for two commonly used exposure factors – body weight (*BW*) and exposure duration (*ED*) – using data for the U.S. population. For these factors we provide a first set of subpopulation based archetypal distributions along with methodology for using these distributions to construct relevant scenario-specific probabilistic exposure factor distributions.

Keywords: Probabilistic Risk Assessment, Probability density function (PDF), Exposure duration, Body weight, Classification and Regression Tree (CART)

INTRODUCTION

Among the current needs of the exposure-assessment community is the need for models and data linking exposure, dose, and health information. Quantifying such links supports environmental surveillance, improves predictive models, and enhances the usefulness of risk assessment and risk management [(1)]. However, high variability and uncertainty in human behaviors, exposure routes, contaminant transport, and human pharmacokinetics requires the risk assessor to collect and analyze large amounts of data and apply sophisticated statistical and mechanistic models to the interpretation of these linkages. This process can be exceedingly difficult and costly to execute. It is also difficult for stakeholders, and governing agency to review.

Practitioners often try to simplify the process by selecting upper bound estimates of uncertain parameters in risk models so that their calculations are accepted as conservative. These upper bound estimates are also transparent for the reviewers. However, compounding upper bound estimates in risk calculations limits opportunities to explore the implications of the risk estimates by identifying margins of error; comparing reducible versus irreducible uncertainties; separating individual variability from true scientific uncertainty, and quantifying benefits, costs, and comparable risks for decision-making. More realistic and sophisticated variance propagation and uncertainty analyses methods are needed to fully explore these source to risk relationships.

The U.S. Environmental Protection Agency (EPA) Office of Emergency and Remedial Response is updating its 1989 Risk Assessment Guidance for Superfund (RAGS) as part of its EPA Superfund reform activities. Volume 3 Part A of RAGS was recently released in draft form as an update to the existing two-volume set of RAGS [(2)]. The update provides policy and technical guidance on conducting probabilistic risk assessment for both human and ecological receptors. However, the effectiveness of these probabilistic methods for characterizing and communicating risk is largely dependent on the risk assessor's ability to characterize the type and degree of uncertainty and variability associated with inputs to their risk models, particularly those inputs that have a strong influence on the modeling outcome.

We believe that the probabilistic risk assessment process would benefit from a systematic approach for developing probabilistic value ranges for model inputs. We envisage an approach that parallels the one adopted by EPA for deterministic risk assessments where the availability of default exposure factor values has improved the ease and consistency of both performing and reviewing deterministic risk assessments. PRAs cannot have an equivalent list of default value ranges for uncertain model inputs. This is primarily because PRAs are applied to confront uncertainty and variability such that assessors can quantify risks at specific percentiles in a cohort or population and set confidence intervals around their predictions. Default input distributions would almost certainly lack the flexibility and necessary relevance for application to specific target populations and exposure scenarios.

We offer instead a standard approach for developing scenario-specific probabilistic inputs. The premise is that homogeneous subsets of the population can be identified for a given exposure factor and archetypal distributions can be developed for these sub-populations. Once the archetypal distributions are developed, one can construct scenario-specific exposure factor distributions without re-collecting or evaluating raw data for each new site-specific risk assessment thereby gaining the benefits of default inputs without loosing the flexibility needed for scenario specific relevance. We present the approach below and demonstrate it's use with two commonly needed exposure factors: body weight and exposure duration.

DESCRIPTION OF THE APPROACH

The approach includes two major steps and each step includes several tasks as summarized in Fig. 1. Step one of the approach is the process of constructing archetypal distributions for relevant demographic subsets of the population. Step two relates these archetypal distributions to site specific characteristics of the population or cohort of interest in a given PRA. Details of each step are provided below.

Step One: Development of Archetypal Distributions for Demographic Subsets

Step one of the approach includes three tasks. First, it is necessary to identify and acquire raw data for the exposure factor of interest along with information about the relevant demographic variables. Second, standard data analysis methods, data mining and graphical techniques are applied to identify the robust demographic variables in the population so the raw data can be partitioned into subsets that are relatively homogeneous with respect to the selected exposure factor. In the final task, archetypal probabilistic distributions are constructed for each subset.

Task 1: Acquiring raw data – Developing archetypal distributions requires population-based data for the exposure factor of interest along with information about the associated demographic variables. When identifying data we consider three key attributes – quantity, quality and relevance. The quantity of raw data is important because a distribution that is based on that data has characteristics (mean, standard deviation, etc) whose uncertainties are inversely related to the sample size. The actual sample size needed to construct a reliable distribution depends on the range of data and the precision in the measurement of each sample value. The quality of data also contributes to uncertainty in the form of experimental error due to flaws in sampling design or imprecision in measurements and/or self-reported information. These errors propagate through to the distribution development process. When evaluating data quality, we generally consider direct measurements of the exposure factor of interest to be best; followed by self-reported values for the exposure factor of interest; direct measures of surrogate data; and finally, self-reported surrogate data. Unfortunately, surrogate information (e.g., estimated breathing rate from caloric intake or exposure duration based on population mobility) is often the only information available. The relevance of data must also be considered because even large amounts of high quality data may not adequately capture temporal and spatial variability or the demographic characteristics of the population of interest. Judging the relevance of data requires a clearly defined exposure scenario, population and/or cohort, well-documented data and a clear description of the steps, if any, taken to adjust or transform the data prior to use [3].

Task 2: Identifying homogeneous subpopulation – A number of exposure factors have been related to demographics such as age and gender [4-8]. However, there are few reported tests of whether these demographic categories are indeed statistically different from one another given the inherent variability in the population and uncertainty in the data. There are a number of approaches that can be used for data analysis when the goal is to partition the data in a way that reduces variance within, and increases the difference between resulting subsets. We apply Classification and Regression Tree (CART) data mining software [9] to systematically identify the most robust demographic variables for a given exposure factor.

CART uses binary recursive partitioning to develop classification or regression trees using a non-analytic, computationally intensive procedure for determining what factor, and value, should be used to split a data set into subsets [9]. Each new subset of the data is then analyzed and split again until either the sample size reaches a lower limit or the cost (added complexity) of an additional split exceeds what would be gained in the form of reduced variance within – and increased distance between – resulting subsets. The technique has been applied in many fields, including engineering, medicine, public health and economics [10-15]. We also use graphical analysis to verify and interpret results from the CART analysis.

Task 3: Constructing archetypal distributions – Once the raw data is acquired and partitioned into relatively homogeneous subsets, the next task is to identify an appropriate statistical model and to parameterize that model to fit the data within each subset. A number of papers, reports and books are available that describe methods for selecting and/or constructing probabilistic input distributions [3,16-21]. In addition, computer software is readily available for automating much of the process. When these methods are applied, one obtains a distribution that provides an optimum fit to the available data.

The approach that we use is to construct an empirical cumulative distribution from the raw data then use exploratory data analysis to identify the simplest (i.e. fewest parameters) statistical model that best captures the range and shape of the empirical distribution. The selected model is then parameterized to fit the data by minimizing the sum of the square difference between the statistical model and the empirical distribution.

Step Two: Developing Scenario-Specific Distributions

At this point we use site-specific information about the size, composition and demographics of the target population to determine how best to sample from the archetypal distributions to construct the most relevant probabilistic input distribution for the exposure factor and population of interest. Population counts are provided in the U.S. Census at scales that include state, county, census tract, block group, block, place, zip code, urbanized area and metropolitan area levels. This information is easily accessed from the Census Bureau's American Fact Finder page [(22)]. The census tables provide enough flexibility to acquire information about a range of demographic variables in the population on a scale that is relevant for most assessments.

After sampling values from the appropriate archetypal distributions in a way that represents the size and demographic composition of the target population, a statistical model is selected and fit to the scenario specific values. When the target population is small, the sampling and distribution fitting process should be repeated several times to get a robust estimate of uncertainty in the statistical parameters of the distribution. This additional information about uncertainty in the statistical parameters can be useful for PRAs that incorporate both variability and uncertainty as discussed by Cullen and Frey [(20)].

ILLUSTRATIVE APPLICATIONS

The following applications illustrate the mechanics and capabilities of the proposed approach with two exposure factors; body weight (*BW*) and exposure duration (*ED*). For these factors we provide preliminary archetypal distributions along with methodology for adapting these distributions to construct a relevant scenario-specific probabilistic exposure factor distribution. Although the following illustrative application used data for the United States, the approach should be applicable to any population given adequate data.

ILLUSTRATIVE APPLICATION 1: BODY WEIGHT

Task 1: Acquisition of Data

Our primary source of data describing body weight (*BW*) for the U.S. population were gathered from the Center for Disease Control revised growth charts for children [23], the Continuing Survey of Food Intakes by Individuals (CSFII) [24] and the most recent National Health and Nutrition Examination Survey (NHANES III) [25].

The revised growth charts are based on a number of national surveys but exclude data from the most recent national health and nutrition survey for children ≥ 6 years to avoid the upward shift in body weights in the U.S. population [23]. Exclusion of this data may reduce the relevance of the *BW* information for the current population particularly if the trend towards increased *BW* continues. We consider the revised growth charts from the CDC to be appropriate for constructing archetypal *BW* distributions for children and adolescents but these distributions will have to be revisited periodically and adjusted to represent changes in the population.

We use information from the 1994 through 1996 CSFII [24] to evaluate and identify the most important demographic variables related to *BW*. After identifying these demographics, the CSFII data for adults (age > 19 years) were combined with NHANES III data to increase the sample size for constructing demographically based *BW* distributions for adults.

Prior to combining the CSFII and NHANES III data, the degree of bias in self-reported *BW* was evaluated by comparing the empirical cumulative distributions for adult males and females from each survey (Fig. 2). The actual measured *BWs* from the NHANES III were assumed to be accurate. We believe that the difference between NHANES III and the CSFII distributions in Fig. 2 is from bias in the self-reported data from CSFII. The jagged nature of the CSFII curves appears to be due to the tendency for self-reported *BWs* to be rounded to the nearest 5 pounds. The overall effect of this bias in the self-reported values was expected to be negligible in the combined data set.

Task 2: Identifying Homogeneous Subsets of the Population

Our analysis of demographic factors for *BW* considered age, gender, race, ethnicity, region, urban (whether individual lived in rural, urban or metropolitan area) and income as reported in Record Type 25 of the CSFII. We excluded respondents that reported being pregnant and/or lactating, resulting in a final sample size of 15,502 persons. The income reported as percent of poverty was converted to categorical values such that "under poverty line"=1, "100-200%of poverty" = 2 and "greater than 200 % of poverty" = 3 (resulting frequency: 1=2673, 2=3667 and 3=9159). We performed the CART analysis set up for a regression tree with v-fold cross validation (n=10) and the minimum cost tree was generated using the least squares method.

The tree diagram in Fig. 3 presents the results for ages 12 years and above. The reported *BW* for children under 12 was primarily dependent on age. For adolescents and adults (age 12 and up) gender became an important variable. Females were separated by age from 12 to 24 years and by race for women 24 years and older. Men were subdivided by age from 12 to 19 years and above 72 and by race for men older than 15 years. The influence of race is further evaluated in Fig. 4 for women showing that the median *BW* reported by "*Asian, Pacific Islander*" women was approximately 12% lower than the general adult female population. The results for men (not shown) were similar with the average reported *BW* by "*Asian, Pacific Islander*" being approximately 20% less than the overall adult male population.

Although race was clearly an important demographic for *BW*, only the CSFII identified respondents as "*Asian, Pacific Islander*" and that data set was not large enough to allow development of separate distributions for each age/gender/race category. Therefore, we constrained this analysis to age and gender categories. If the target population in a PRA is composed predominantly of adults of Asian and/or Pacific Islander decent, then we recommend that the risk assessor re-evaluate the raw data or collect new data to develop relevant distributions for the specific cohort.

Task 3: Constructing Archetypal Distributions for BW based on Age and Gender

We used the information described above to develop *BW* distributions for each age/gender category. The categories were selected to coincide with the available information from the U.S. Census summary tables. These tables include yearly values by gender from birth through 19 years (Census Table P14) and 18 additional age categories for adults (Census Table P12).

Children (birth through 19 years) – For children from birth through 19 years, we begin with the statistically smoothed growth curves from the CDC [(23)] and transform them into probability distributions that match the age/gender categories reported in the census summary tables. To do this, we used the growth curves from the original CDC report and simulated 2500 *BW* values for each 1-month age interval. We then grouped the data to match the age/gender categories reported in Census Table P14. For example, we used the 2500 body weight values for each 1-month interval from 24 to 36 months (n=30,000) to construct the archetypal distribution of *BW* for 2 year olds. We use exploratory data analysis to determine that the 3-parameter lognormal model [(26)] works well for all age categories. We fit the 3-parameter lognormal to the data from the growth charts by minimizing the sum of the square differences (SS_{xy}) between the empirical and estimated percentiles. This "fit" was performed on the 1^{st}, 5^{th}, 10^{th}, 15^{th}, …, 85^{th}, 90^{th}, 95^{th} and 99^{th} percentiles of the data. The resulting parameters for the archetypal *BW* distributions for children from birth through 19 are reported in Table 1.

Adults (≥ 20 years) – For adults, we started with raw *BW* data from the combined CSFII and NHANES III. We grouped the data for adults by gender and the following ages (years) – 20, 21, 22 through 24, 5 year intervals from 25 through 59, 60 and 61, 62 through 64, 65 and 66, 67 through 69, 5 year intervals from 70 through 84 and 85 and above. This gave a total of 18 age categories for each gender that corresponded to data reported in Census Table P12. Once the data were grouped, we estimated the empirical percentiles for each age/gender category. We then smoothed these raw percentiles using the functions

$$y = a + bx^3 + \frac{c}{\sqrt{x}} \tag{1}$$

for males and

$$y = a + bx^2 + cx^{2.5} \tag{2}$$

for females, where x is age (years) at the midpoint of the category and y is the *BW* at the given percentile. The parameters *a, b,* and *c* are provided in Table 2 and can be used to estimate percentiles for age categories that are not included here. Both the raw percentiles and the smoothed percentiles for each age category are provided elsewhere [(27)].

Fig. 5 plots the 5th, 50th and 95th percentiles of *BW* and the empirical values for adult males as a function of age along with curves for children (birth through 19 years). The slight discontinuity at the transition from the children to the adult curves, particularly for the 95th percentile, is likely due to the exclusion of the NHANES III data from the CDC growth charts (e.g., omission of the effect of overweight persons in recent surveys). The trend in Fig. 5 is similar to one seen in cross-sectional data from the NHANES II survey [(28)] and to the longitudinal long-term variation in an individual's BW reported by Okajima [(29)].

As with the children's data, the exploratory analysis revealed that the 3-parameter lognormal was adequate for fitting the *BW* values within the different age/gender categories. Therefore, we used the same method previously described for constructing *BW* distributions for children to fit the 3-parameter lognormal model for adults. Table 3 provides the resulting parameters for the archetypal *BW* distributions for adults in different age/gender categories.

Step 2: Constructing Scenario-specific Distributions of BW

Summary tables from the 2000 U.S. census data provide information on the demographic composition of the population at different geographically relevant scales for the U.S. population [22]. Census Table P14 provides 100% counts by gender for each year from birth through 19 and Census Table P12 give an additional 18 age categories for adults twenty years and older. These age and gender specific counts form the basis for constructing *BW* distributions using the archetypal distributions described earlier.

Based on data from the Census tables an appropriate number of individual *BW* values are drawn from each archetypal age/gender specific distribution creating a single realization of *BWs* for the population or cohort of interest. The *BW* values can be drawn and grouped to address specific needs of the PRA. (e.g., children ages 7-13 or adult women living within a particular exposure district). Finally, a statistical model is selected and fit to the simulated data. For *BW*, typically a 2- or 3-parameter lognormal is appropriate. When the target population is relatively small, we determined that the sampling and fitting process should be repeated several times to characterize uncertainty about the choice of statistical model and parameters.

ILLUSTRATIVE APPLICATION 2: EXPOSURE DURATION

Defining Exposure Duration

Unlike *BW*, exposure duration (*ED*) has alternate definitions. Some of the more common definitions include time spent in a given location, occupation, or participating in a particular activity. In general, exposure duration is the time interval during which exposure occurs, either continuously or intermittently, at a given exposure concentration and intake/uptake rate [(30, page 1-12)]. For this illustration we define *ED* as the time spent in an "exposure district" (i.e., the area that is potentially impacted by a particular harmful agent and/or contaminated site).

Task 1: Sources of Data

A useful first approximation for *ED* as defined above is the amount of time an individual remains in his/her current residence – previously referred to as "total residence time" [(31)], "residential occupancy period" [(4)] and "residence duration" [(32)]. However, approximately half of homebuyers purchase homes within 10 miles of their previous residence [(30, Table 15-171)]. Thus, changing residence at the end of an occupancy period does not necessarily mean that the exposed individual has moved out of harm's way. Therefore, the estimate of *ED*, as defined in this application, requires knowledge of both the total occupancy period and the likelihood that the end of the occupancy period results in a move out of the exposure district.

Nationally representative data on total occupancy period are not available. Rather, surrogate data such as mobility, mortality, current residence time and/or tax records have been used to estimate occupancy period [(31, 4, 33, 32)]. Information on distance of move is also lacking. The most readily available surrogate for distance of move is from migration data that reports movement on the scale of state, county, metropolitan area, central city, or suburb.

The U.S. Bureau of the Census conducts the American Housing Survey (AHS) [22] providing comprehensive housing statistics for the U.S. Department of Housing and Urban Development. The surveys include information on housing type, attributes of housing units, and data on household members. The Census also provides information on population mobility [34], and identifies movers and non-movers by their reported change in place of residence from one survey to the next. The Census further classifies movers as to whether they relocated to the same or different county, state, or region; and whether they are movers from abroad, within or between central cities, suburbs, or non-metropolitan areas. We used the housing survey data (1995 survey where n=58,318 households) to calculate current residence time (CRT) based on reported "Year householder moved into unit". We used the calculated CRT as a surrogate for occupancy period to identify robust demographic variables for *ED*. We then used mobility data to develop archetypal and scenario specific distributions.

Task 2: Identifying Homogeneous Subsets of the Population for ED

The demographic variables in the CART analysis of *ED* included tenure (whether owner or renter occupied), age (0 –90 by year and >91yrs), gender, race, Spanish origin, household income as percent of poverty, and census region. We also considered whether the household was located in a central city or suburban area and if the residence was a farming household. The AHS 1995 variable 'crop sales' (the sales of agricultural products from the farm units were $1,000 or more in the last 12 months prior to the interview) was used to identify farming households.

The classification tree in Fig. 6 shows that differences in reported CRT were primarily dependent on age and tenure. There was also a small regional difference for people older than 68 years where those residing in the Northeast had average reported CRT values about 5 years longer than the general population. Fig. 7 shows that further evaluation of the age dependence of CRT revealed a bimodal relationship that peaks near age 18 then dropped to a minimum at around age 30 followed by a steady increase. This bimodal distribution may be due to a direct correlation between the mobility of children and their parents. If we exclude children, the relationship between residence time and age is approximately linear. Migration data, also reported in the census, indicate that the median occupancy period remains relatively constant across age categories in owner occupied homes but doubles between ages 24 and 65 for renters [35]. Furthermore, renters are approximately 4 times more likely to move in a given year than owners and those that do move are approximately 3 times more likely to move to a different county. Therefore, although age was related to CRT, we concluded that the most important information for characterizing the likelihood and distance of a move was the composition of the housing stock in the exposure district (i.e., rental versus owner occupied).

Task 3: Constructing Archetypal Distributions for ED based on Tenure

In constructing the *ED* distributions, we assumed no prior knowledge of individual ages within each home in an exposure district. Rather, we estimated archetypal distributions of *ED* in terms of occupancy period for a given household using tenure-based mobility data. Mobility data gives the likelihood that an individual within a certain demographic category will move during the year prior to a survey. For total movers and movers out of a county, the annual moving rates in the general population are relatively constant over time as shown in Fig. 8. A population having a constant rate of change can be described using a probabilistic model from the family of "life distribution models" [26], the simplest of which is the exponential model, which is parameterized using a single scale parameter such that the probability of moving in a given year is given by $1-\exp(-x/b)$ where x is the year and b is the scale parameter.

To parameterize archetypal distributions based on tenure, we used data from 1989-1999 and estimate the average moving rate for owner- and renter-occupied households along with the moving rate to a different county. Assuming that the annual variation in mobility rate is normally distributed, we obtained for owner occupied households annual average moving-rate scale parameters (± 1 standard deviation) of 0.085 ± 0.004 y^{-1} and 0.324 ± 0.009 y^{-1} for total moves and moves out of county, respectively. For renter occupied households the moving-rate scale parameters were 0.033 ± 0.003 y^{-1} and 0.112 ± 0.007 y^{-1} for, respectively, total moves and moves out of county. Thus, the archetypal distributions for *ED* were simply rate constants specified for each demographic subset of the population.

Step 2: Constructing Scenario-specific Distributions of ED

Using the moving rates described above one can calculate the *ED* for a given household in the housing stock as

$$ED_i = \sum_{i=1}^{n_i} Y_i \qquad (3)$$

where

$$Y_i = \begin{cases} 1 \text{ year} & \text{if } R < P_d \\ 0 & \text{if } R \geq P_d \end{cases} \qquad (4)$$

and R is a random number drawn from the unit rectangular variate, P_d is the tenure-based probability of moving as derived from the exponential model and n_i is the total number of sequential random draws that satisfy the constraint $R < P_d$. This process is repeated for a given home over a defined averaging time (*AT*) generating a profile of occupancy periods for that residence.

We extend Eq. 3 to the total number of homes in an exposure district by determining the appropriate number of renter- and owner-occupied households in the exposure district and constructing a relevant sample of *ED* values over the specified *AT*. The sampling process using Equations 3 and 4 is easily applied using available statistical, mathematical or spreadsheet software packages. The risk assessor can use a sample of *ED* values generated in this way to construct the probabilistic distribution for the scenario of interest using standard methods for identifying appropriate distributions and fitting them to data.

Information about the composition of the housing stock within an exposure district can be accessed through the U.S. Census or by direct survey of the target population. Specifically, the mobility of different demographic subsets of the population is available at http://www.census.gov. In applying this method to different geographical regions of the U.S., we found that the general exponential distribution including both location and scale [36] consistently provided the best fit to different sets of simulated exposure duration values. For exposure districts where the number of houses is small, we recommend that the sampling and fitting process be repeated several times to estimate uncertainty about model parameterization. This approach can easily be extended to include other demographic characteristics of a target population but given the generality used for distance of move (i.e., out of county), it is not clear whether the added complexity is warranted.

DISCUSSION

Risk assessors use exposure models to describe the relative magnitude and variation in human contact with environmental contaminants. An important measure of an exposure model's performance is its ability to account for those factors that explain variation in this contact (i.e. age, gender, location, activity patterns and physiological characteristics). Uncertainties limit the ability of models to fully characterize these relationships. The uncertainty and variability that is contributed to exposure/risk models by model inputs can be addressed in part through the use of probabilistic input distributions in the context of a PRA. But the process can become haphazard and inconsistent without a standard approach for developing scenario-specific input distributions in PRA. It is critical that the standard approach be consistent but also maintain enough flexibility to address the range of exposure and risk scenarios that might exist.

We present here an approach that provides default or archetypal distributions for exposure parameters in well defined subsets of the population and a process for adapting these archetypal distributions to specific scenarios and populations. This approach has two steps. The first step provides information that is general to most PRAs in the form of archetypal distributions for subsets of the population, similar to the default inputs used in deterministic risk assessment. The second step provides the method to adapt these defaults to scenario specific conditions, resulting in a probabilistic input distribution that is relevant to the target population. We provided two examples to illustrate this two-step approach—one with body weight (*BW*) and one with exposure duration (*ED*).

The first application using *BW* illustrates a case where a lot of highly relevant and precise data is available on a national scale, including significant detail about demographics. Even so, we found that these data still lack the information needed to develop archetypal distributions for all of the important demographic subsets of the U.S. population. Using both a CART and graphical analysis, we identified race as an important demographic variable for *BW*, but found insufficient data to construct archetypal distributions for all relevant combinations of age, gender and race. Future studies could address this shortcoming by incorporating more subjective techniques for defining archetypal distributions for race, particularly for individuals of "*asian/pacific islander*" descent. Nevertheless, the archetypal distributions that we developed provide a first example of an approach that incorporates the benefits of "default inputs" into the PRA process without sacrificing the flexibility that is necessary for relating the analysis to specific target populations and/or risk scenarios.

In contrast to the *BW* application, the *ED* application illustrated a case where actual measurements of the value of interest are virtually nonexistent but where a preponderance of predictive tools using surrogate data are available. For example, Israeli and Nelson, [31] used data from the 1985 and 1987 American Housing Surveys to estimate expected total residence time for a number of demographic categories. They fit a five-parameter survival function to mobility data then used three of the five fitted parameters to estimate the expected total residence time. Johnson and Capel [4] used a Monte Carlo approach to develop distributions of residential occupancy time by gender and age based on both mobility and mortality tables. Finley *et al.* [37] summarized the work of both Israeli and Nelson, and Johnson and Capel then developed additional residential occupancy periods of children born in a given household based on moving rates. Price et al. [33] use a simulation approach with mortality and mobility data to estimate what they term the "Fraction of Life Exposed", which is the exposure duration divided by the lifetime of a hypothetical individual. The modeled person is generated to represent current distributions of age and gender in the U.S. population then a lifetime of exposure for that individual is simulated based on age specific mobility and mortality data. Finally, Sedman et al. [32] used public records on property title transfers as a surrogate for occupancy period of owner occupied households. The advantage of the Sedman et al. method was that the records provide a closed interval for total residence duration.

All of these existing methods for estimating *ED* gave comparable results as illustrated in Fig. 9. Results from the archetypal tenure-based distributions using a nationally representative ratio of renter/owner households and moving rates out of county were in good agreement with existing models (see Fig. 9). The novel element of the approach described here is the identification of important demographic variables that account for relatively homogeneous subsets of the population and the use of these subsets to develop archetypal probabilistic distributions that can be used to construct distributions for a range of different populations and scenarios. A key advantage of this approach is that it reduces the tendency for risk professionals to develop models that go beyond what data and theory can support. For example, given the lack of precision in defining an exposure district and the lack of detailed data on distance of move, it is difficult to justify a more complicated model of *ED.* Another advantage of the standardized approach described here is that it is expected to make assumptions in the PRA process more transparent thereby easing the burden on risk managers and government agencies who are charged with reviewing and using results from the PRA.

The standard method and archetypal distributions described here can benefit from a systematic and critical evaluation against an independent set of well-characterized data for a specific population. In addition, identification of homogeneous subsets and archetypal distributions for other exposure factors is still needed.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Environmental Protection Agency and carried out at Lawrence Berkeley National Laboratory through the U.S. Department of Energy under Contract Grant No. DE-AC03-76SF00098. Environmental Protection Agency funding was provided by the Office of Emergency and Remedial Response (OERR) through Interagency Agreement #DW-899-38067-01-0 and by the National Exposure Research Laboratory through Interagency Agreement # DW-988-38190-01-0

REFERENCES

1. **NAS.** 1994. Science and Judgment in Risk Assessment. Book. National Research Council, Commission on Life Sciences, Board on Environmental Studies and Toxicology, Committee on Risk Assessment of Hazardous Air Pollutants.
2. **USEPA.** 1999. Risk Assessment Guidance for Suprefund (RAGS): Volume 3A Process for Conducting Probabilistic Risk Assessment. DRAFT EPA 000-0-99-000. Office of Emergency and Remedial Response.
3. **USEPA.** 1999. Report of the Workshop on Selecting Input Distribution for Probabilistic Risk Assessment EPA/630/R-98/004. U.S. Environmental Protection Agency, Risk Assessment Forum.
4. **Johnson, T., and J. Capel.** 1992. A Monte Carlo Approach to Simulating Residential Occupancy Periods and Its Application to the General U.S. Population. Report EPA 450/3-92-011. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Emissions Standards Division.
5. **Burmaster, D. E.** 1998. Lognormal Distributions for Skin Area as a Function of Body Weight. Risk Analysis **18:**27-32.
6. **Burmaster, D. E., and D. M. Murray.** 1998. A Trivariate Distribution for Height, Weight, and Fat of Adult Men. Risk Analysis **18:**385-389.
7. **ODEQ.** 1998. Guidance for Use of Probabilistic Analysis in Human Health Risk Assessments (Interim Final). Oregon Department of Environmental Quality, Waste Management and Cleanup Division.

8. **USEPA.** 2000. Estimated Per Capita Water Ingestion in the United States Based on Data Collected by the United States Department of Agriculture's 1994-1996 Continuing Survey of Food Intakes by Individuals. Report EPA-822-R-00-008. U.S. Environmental Protection Agency, Office of Water.

9. **Breiman, L., J. Friedman, R. Olshen, and C. Stone.** 1984. Classification and Regression Trees. Wadsworth, Pacific Grove.

10. **Spear, R. C., F. Y. Bois, T. Woodruff, D. Auslander, J. Parker, and S. Selvin.** 1991. Modeling Benzene Pharmacokinetics Across 3 Sets of Animal Data - Parametric Sensitivity and Risk Implications. Risk Analysis **11:**641-654.

11. **Spear, R. C., T. M. Grieb, and N. Shang.** 1994. Parameter Uncertainty and Interaction in Complex Environmental Models. Water Resources Research **30:**3159-3169.

12. **Eisenberg, J. N., E. Y. W. Seto, A. W. Olivieri, and R. C. Spear.** 1996. Quantifying Water Pathogen Risk in an Epidemiological Framework. Risk Analysis **16:**549-563.

13. **Pilote, L., D. P. Miller, R. M. Califf, J. S. Rao, W. D. Weaver, and E. J. Topol.** 1996. Determinants of the Use of Coronary Angiography and Revascularization After Thrombolysis For Acute Myocardial Infarction. New England Journal of Medicine **335:**1198-1205.

14. **Eisenberg, J., D. H. Bennett, and T. E. McKone.** 1998. Chemical dynamics of persistent organic pollutants: A sensitivity analysis relating soil concentration levels to atmospheric emissions. Environmental Science & Technology **32:**115-123.

15. **Eisenberg, J., and T. E. McKone.** 1998. Decision Tree Method for the Classification of Chemical Pollutants: Incorporation of Across-Chemical Variability and Within-Chemical Uncertainty. Environmental Science and Technology **32:**3396-3404.

16. **D'Agostino, R. B., and M. A. Stephens (ed.).** 1986. Goodness of Fit Techniques, vol. 68. Marcel Dekker, Inc., New York.

17. **USEPA.** 1997. Guiding Principles for Monte Carlo Analysis. Report EPA/630/R-97/001. Risk Assessment Forum, U.S. Environmental Protection Agency.

18. **USEPA.** 1998. Guidance for Submission of Probabilistic Exposure Assessments to the Office of Pesticide Programs. Draft. U.S. Environmental Protection Agency, Office of Pesticide Programs.

19. **Conover, W. J.** 1999. Practical Nonparametric Statistics, 3rd ed. John Wiley & Sons, New York.

20. **Cullen, A. C., and H. C. Frey.** 1999. Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs. Plenum Press, New York.

21. **USEPA.** 2000. Options for Development of Parametric Probability Distributions for Exposure Factors EPA/600/R-00/058. U.S. Environmental Protection Agency, National Center for Environmental Assessment, Office of Research and Development.

22. **Census, U. S.** United States Census 2000. U.S. Census Bureau. http://www.census.gov/ (last accessed February 20, 2003)

23. **Kuczmarski, R. J., C. L. Ogden, L. M. Grummer-Strawn, and a. others.** 2000. CDC Growth Charts: United States, Advanced data from vital and health statistics. Revised no. 314. National Center for Health Statistics.

24. **USDA.** 1998. 1994-96 Continuing Survey of Food Intakes by Individuals and 1994-96 Diet and Health Knowledge Survey Accession No. PB98-500457. Department of Agriculture, Agricultural Research Service.

25. **NCHS.** 1996. National Health and Nutrition Examination Survey, III (NHANES III) 1988-94 reference manuals and reports (CD-ROM). CD-ROM Series 11 No. 1. National Center for Health Statistics, Department of Health and Human Services, Centers for Disease Control and Prevention.

26. **NIST/SEMATECH.** e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, (last accessed February 21, 2003)

27. **Maddalena, R. L., and T. E. McKone.** 2002. Developing and Evaluating Distributions for Probabilistic Human Exposure Assessments. Report LBNL-51492. Lawrence Berkeley National Laboratory.

28. **Burmaster, D. E., and E. A. C. Crouch.** 1997. Lognormal Distributions for Body Weight as a Function of Age For Males and Females in the United States, 1976-1980. Risk Analysis **17:**499-505.

29. **Okajima, Y., M. Togo, G. Kitagawa, and S. Nishikawa.** 2000. Time Series Analysis of Monthly Body Weight and Blood Pressure of One Man from 29 to 65 years. American Journal of Human Biology **12:**526-541.

30. **USEPA.** 1997. Exposure Factors Handbook (Final) Volumes I, II, III EPA/600/P-95/002Fa. U.S. Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment.

31. **Israeli, M., and C. B. Nelson.** 1992. Distribution and Expected Time of Residence for U.S. Household. Risk Analysis **12:**65-72.

32. **Sedman, R., L. M. Funk, and R. Fountain.** 1998. Distribution of Residence Duration in Owner Occupied Housing. Journal of Exposure Analysis and Environmental Epidemiology **8:**51-57.

33. **Price, P. S., J. Sample, and R. Strieter.** 1992. Determination of Less-Than- Lifetime Exposures to Point Source Emissions. Risk Analysis **12:**367-382.

34. **USDOC** 1999, Annual Geographical Mobility Rates, By Type of Movement: 1947-1999 Table A-1. U.S. Census Bureau. http://www.census.gov/ (last accessed February 20, 2003)

35. **Hansen, K. A.** 1998. Seasonality of Moves and Duration of Residence P70-66. U.S. Bureau of the Census, Current Population Reports.

36. **Evans, M., N. Hastings, and B. Peacock.** 1993. Statistical Distributions, Second Edition ed. John Wiley & Sons, Inc., New York.

37. **Johnson, N. L., S. Kotz, and N. Balakrishnan.** 1994. Continuous Univariate Distributions, Second ed, vol. 2. John Wiley & Sons Inc., New York.

38. **Finley, B., D. Proctor, P. Harrington, N. Paustenbach, and P. Price.** 1994. Recommended Distributions for Exposure Factors Frequently Used in Health Risk Assessment. Risk Analysis **14:**533-553.

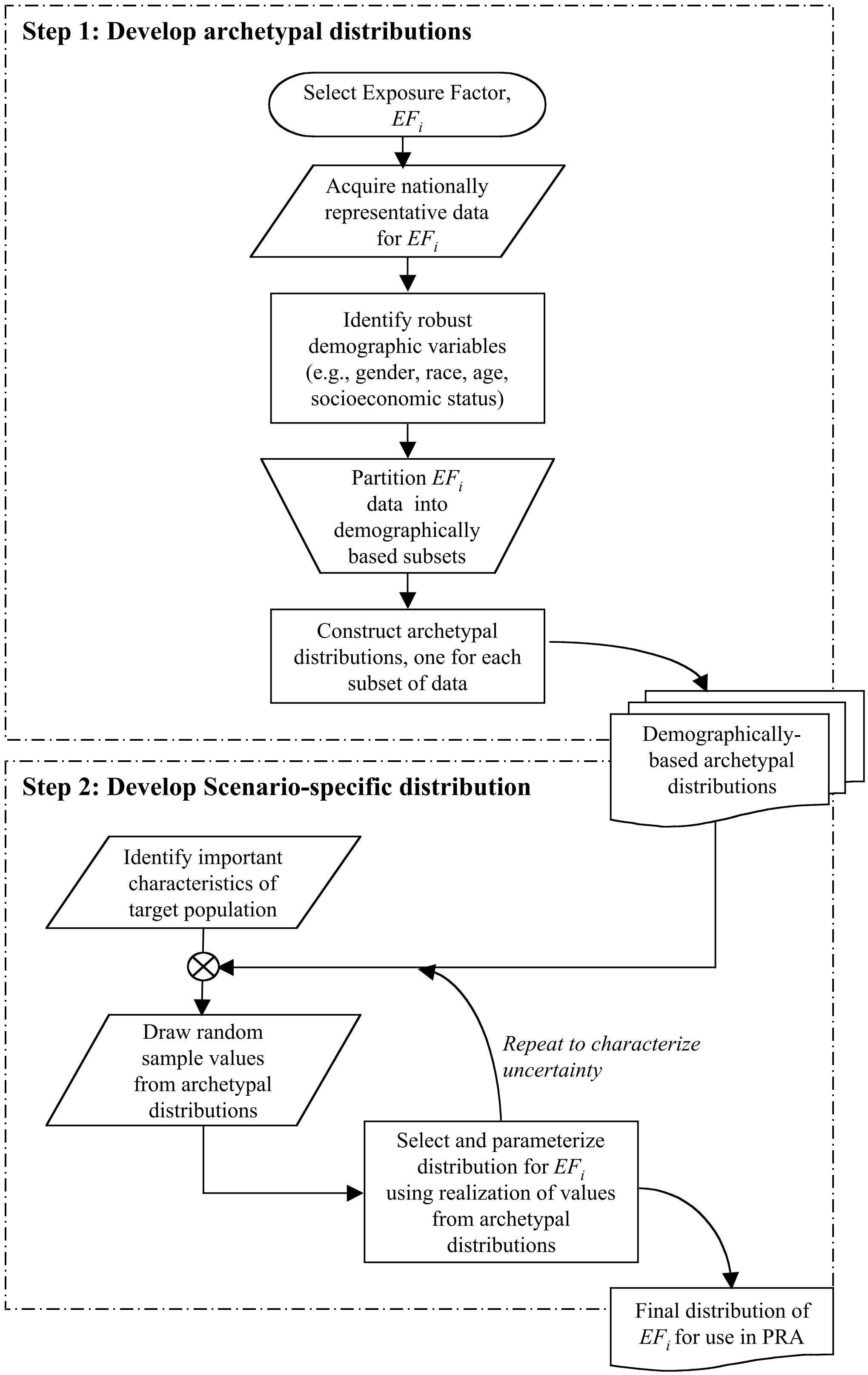


Fig. 1: Schematic representation of standardized approach for developing exposure factor distributions

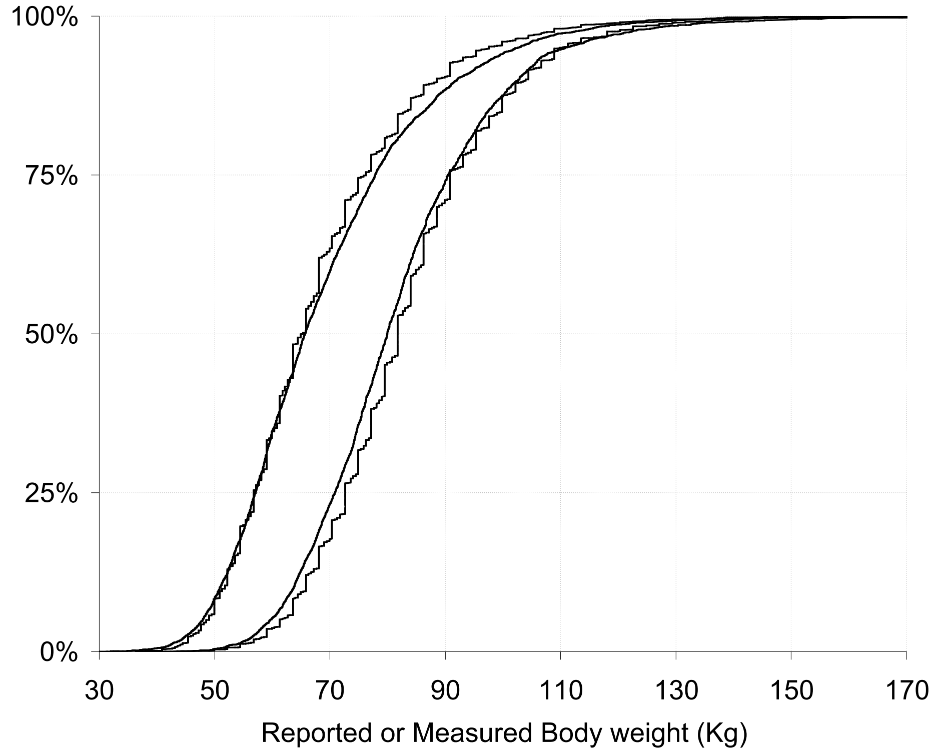


Fig. 2: Comparison of measured *BW* values from NHANES III (smooth curves) and self-reported *BW* values from CSFII (irregular or jagged curves).

CART Output for Body Weight (all sample persons 12 y and older)

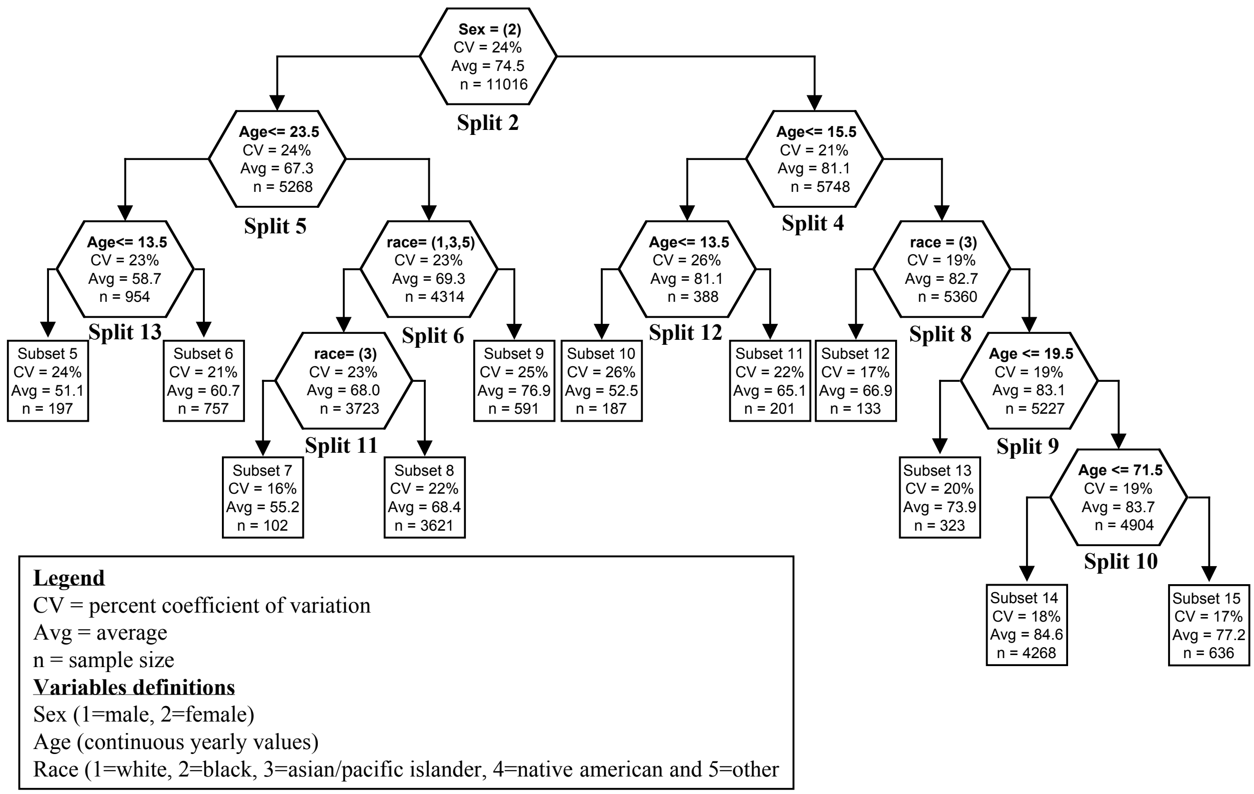


Fig. 3: Classification and regression tree from CART analysis of the CSFII *BW* (kg) data. The tree begins at the second data split because *BW* of children under age 12 is dependent primarily on age. The figure is read as a binary decision tree. A logical statement is given at each decision point (hexagon) (i.e., "Sex" = (2)). Data for which the statement is true move to the left creating a new data subset. When the statement is false, the data move to the right. The average, coefficient of variation and sample size is reported for each node. The "split" order indicates the relative importance of that split in decomposing the data. Terminal nodes that are not shown include respondents age 1 and 2, 3 – 6, 7 – 9 and 10 – 11 years.

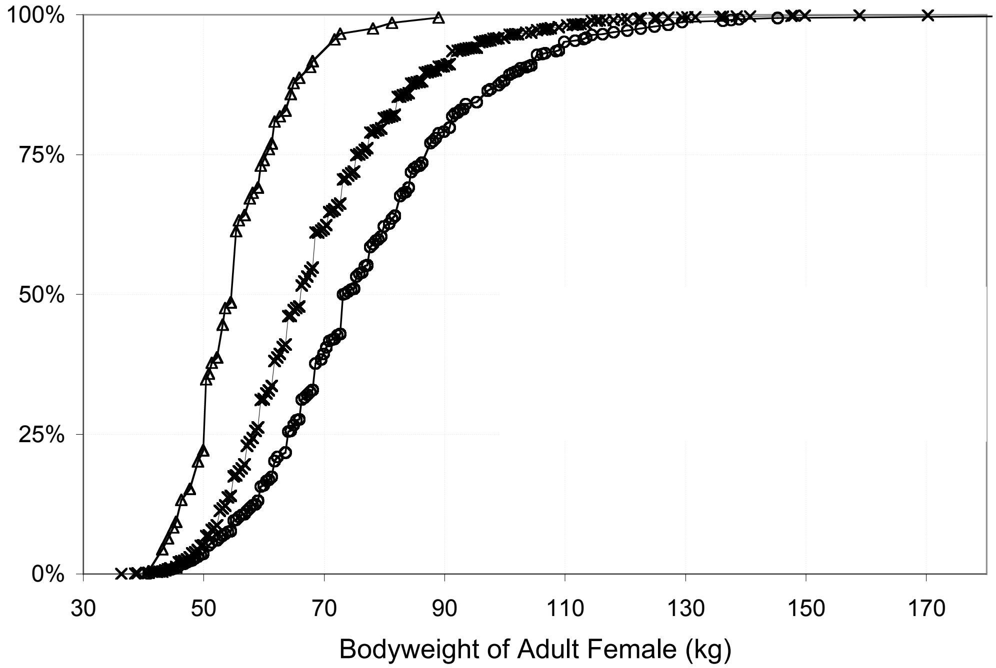


Fig. 4: Importance of race on reported *BW* from CSFII for females. Similar pattern was found for the reported body weight for adult men (data not shown).

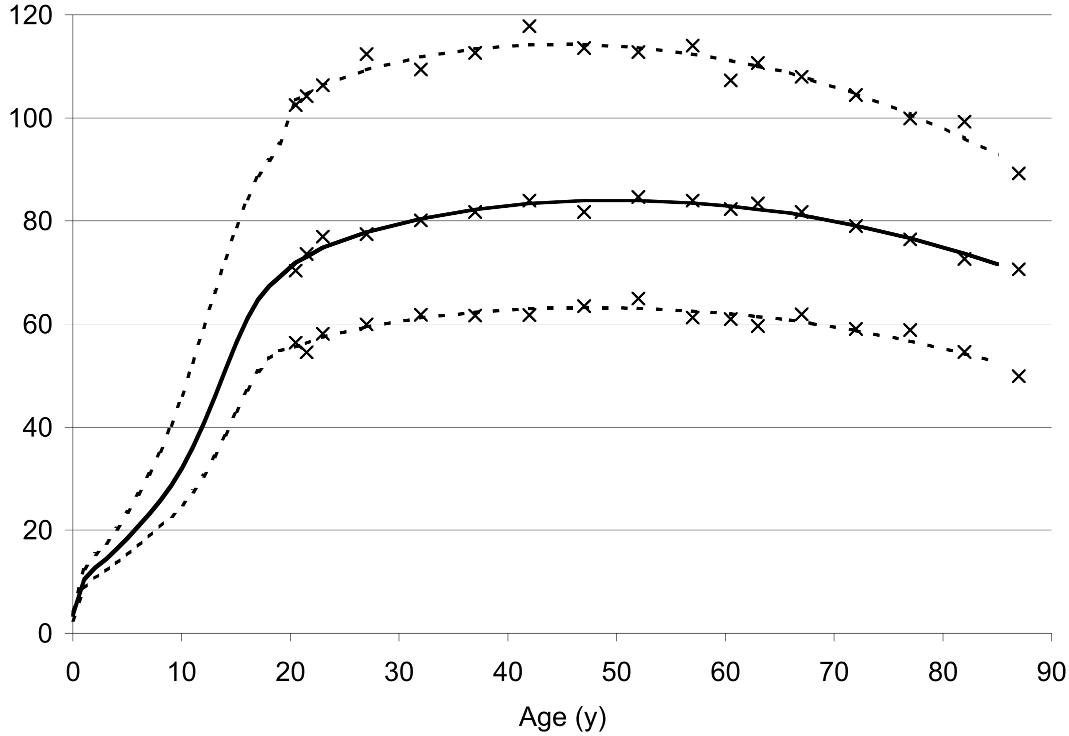
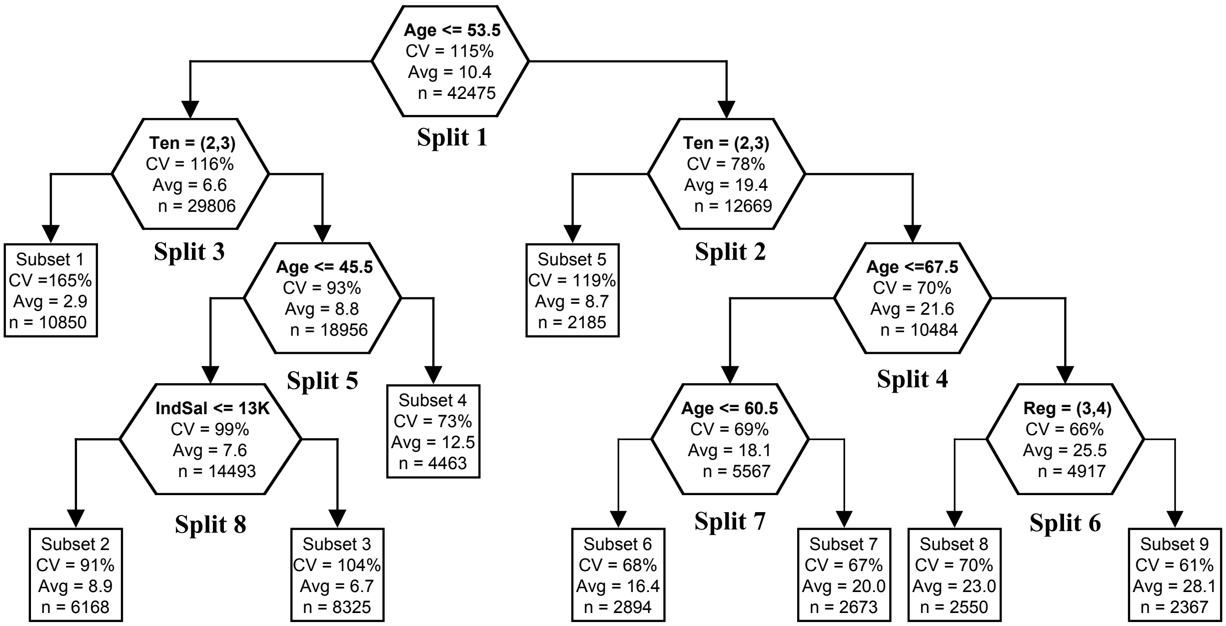


Fig. 5: Illustration of the 5th (lower dash), 50th (solid line) and 95th (upper dash) percentiles of *BW* for adult males as a function of age. Curves to the left of the solid vertical line are based on the CDC growth charts and those to the right are from the combined CSFII, NHANES III data. The marks (□) represent empirical percentiles of the raw data prior to smoothing for each age category.

CART Output for reported Current Residence Time (Sample persons age > 18 years)



Legend
CV = percent coefficient of variation
Avg = average
n = sample size
Variables definitions
Age (continuous yearly values)
Reg = Region (1=northeast, 2=midwest, 3=south, 4= west)
Ten = Tenure (1=owner occupied, 2=rental unit, 3=no cash rent)
IndSal = Annual Salary of Individual in dollars (continuous 0 - 100K, categorical > 100K)

Fig. 6: Classification and regression tree from CART analysis of current residence time (y) data. The tree excludes sample persons younger than 18 years (see text). See Fig. 3 for instructions on how to read tree.

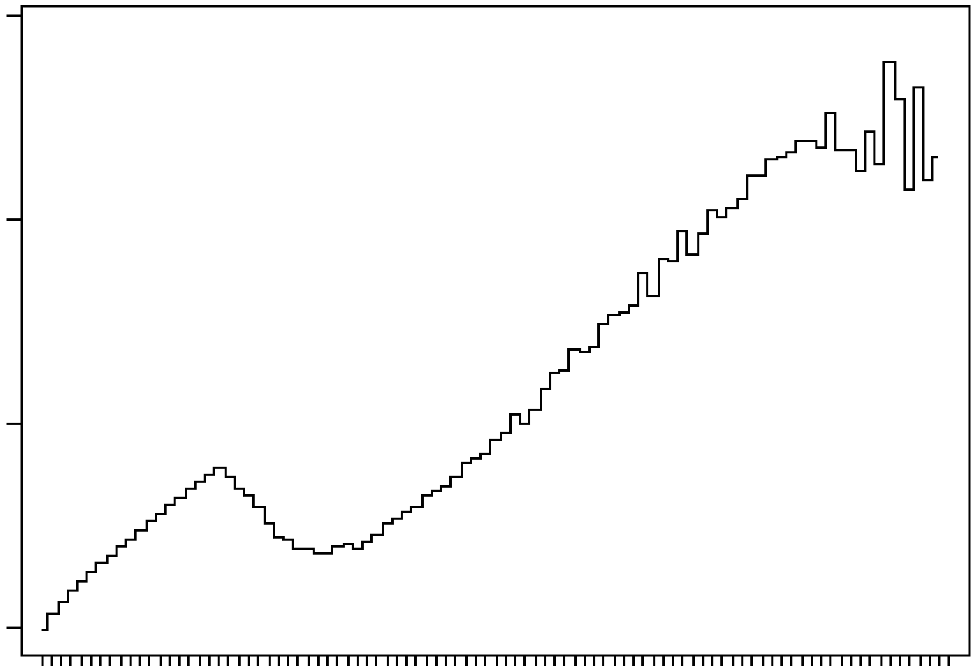


Fig. 7: The relationship between the average reported CRT for the national population and the age at interview showing a bimodal distribution that may be due in part to the expected correlation between CRT of children and their parents.

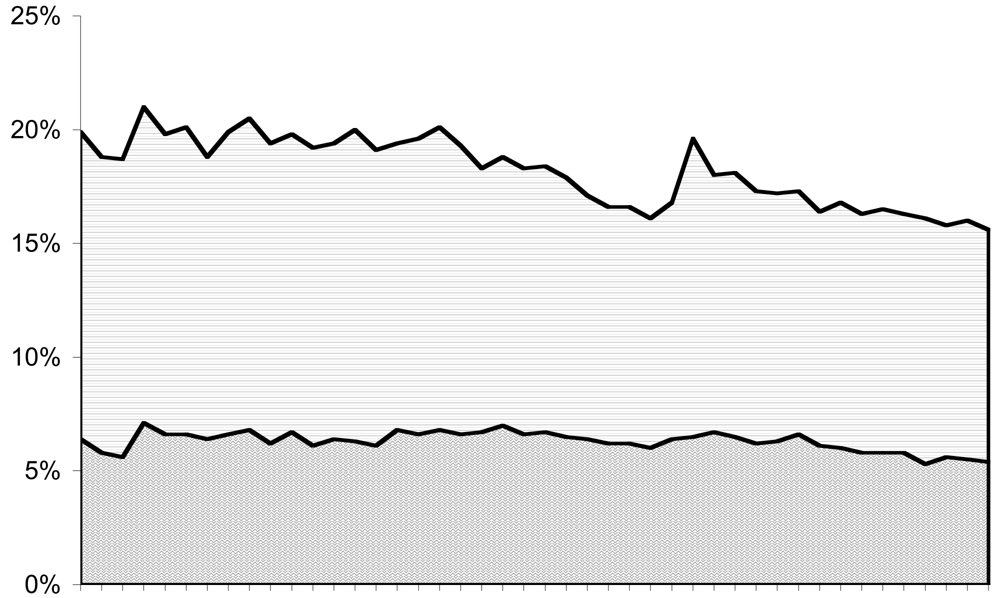


Fig. 8: Annual geographic mobility data showing percent of population that reported moving during the year that proceeded each Annual Demographic Supplement to the Current Population Survey. The data used to generate this figure are from the U.S. Census [22].

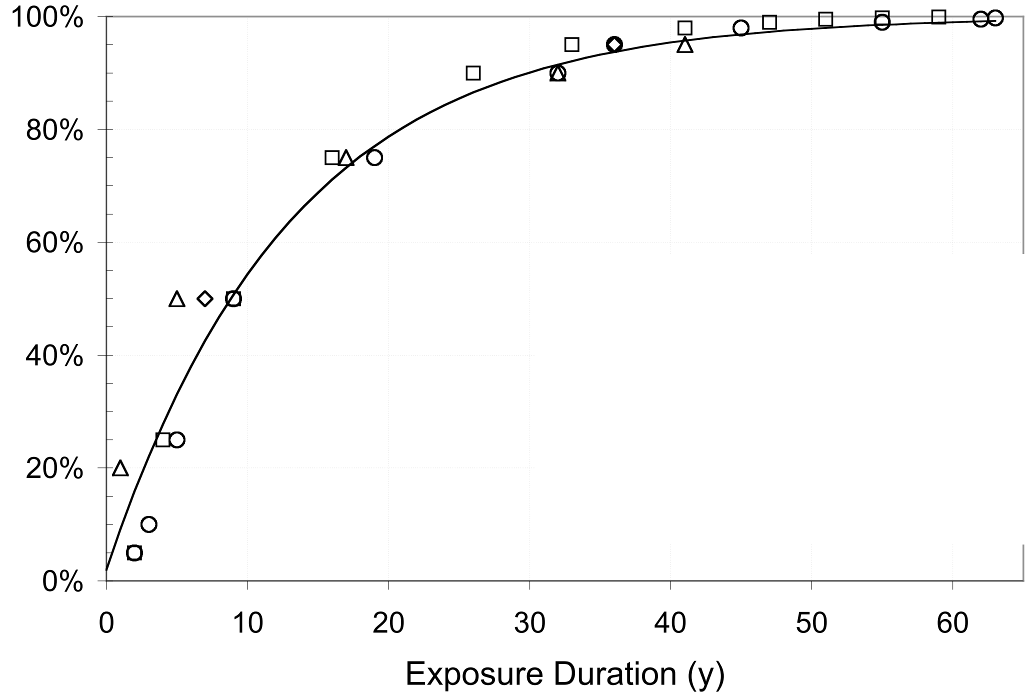


Fig. 9: Comparison of the estimates for occupancy period (total time that an individual lives in current residence) from different methods. The solid line is derived from frequency (used here as probability) of moving to a different county during a given year as described in the text.

Table 1: Parameters for the 3-parameter lognormal archetypal *BW* distributions for each age/gender category for children

Age category (years)	Males				Females			
	γ	$\mu_{LN(x)}$	$\sigma_{LN(x)}$	SS_{xy}	γ	$\mu_{LN(x)}$	$\sigma_{LN(x)}$	SS_{xy}
0	-39.714	3.858	0.050	9.05E-03	-33.487	3.701	0.053	7.74E-03
1	0.499	2.409	0.130	2.09E-05	0.810	2.314	0.136	1.49E-05
2	3.449	2.306	0.156	3.66E-05	5.463	2.013	0.208	5.98E-05
3	6.595	2.155	0.213	2.50E-05	7.743	1.954	0.276	3.52E-05
4	8.398	2.185	0.252	5.67E-05	9.092	2.047	0.310	3.56E-05
5	9.783	2.277	0.276	1.25E-04	10.586	2.135	0.336	4.41E-05
6	11.228	2.362	0.303	5.87E-05	11.592	2.292	0.346	7.48E-05
7	12.978	2.430	0.334	4.06E-05	12.885	2.416	0.369	5.60E-05
8	15.592	2.439	0.393	1.17E-04	13.528	2.618	0.365	1.25E-04
9	17.119	2.565	0.413	1.38E-04	13.841	2.840	0.356	9.29E-05
10	17.857	2.773	0.403	6.05E-05	15.601	2.964	0.371	6.66E-05
11	17.994	3.003	0.376	5.74E-05	16.407	3.136	0.355	8.82E-05
12	18.013	3.215	0.346	7.31E-05	18.746	3.220	0.356	8.34E-05
13	18.494	3.394	0.320	4.57E-05	23.016	3.209	0.383	1.26E-04
14	19.736	3.523	0.301	3.61E-05	26.543	3.189	0.388	1.58E-04
15	22.236	3.595	0.290	1.19E-04	30.961	3.091	0.423	2.18E-04
16	29.143	3.522	0.322	3.44E-05	34.174	3.012	0.450	2.96E-04
17	31.499	3.540	0.316	5.66E-05	35.942	2.982	0.463	2.13E-04
18	34.847	3.505	0.332	8.30E-05	36.063	3.029	0.454	2.25E-04
19	34.932	3.553	0.326	4.57E-05	35.325	3.119	0.434	2.59E-04

As an example, *BW* values for a given age/gender category can be calculated using Microsoft Excel functions as BW_i = (LOGINV(RAND( ), $\mu_{LN(x)}$, $\sigma_{LN(x)}$))+ γ where γ is the shift or location parameter, $\mu_{LN(x)}$ and $\sigma_{LN(x)}$ are the arithmetic mean and standard deviation of the distribution of ln(x) and the function RAND( ) generates a random number from the unit rectangular variate. Note that the negative offset parameter for children birth to 1 y will occasionally result in a negative value, which should be discarded prior to constructing a final distribution.

Table 2: Values for estimating the smoothed age specific percentiles of *BW* for adults

	Males (Equation 1)			Females (Equation 2)		
Percentile	*a*	*b*	*c*	*a*	*b*	*c*
0.01	7.31E+01	-2.71E-05	-9.80E+01	3.93E+01	1.16E-02	-1.33E-03
0.05	8.64E+01	-3.08E-05	-1.37E+02	4.34E+01	1.32E-02	-1.46E-03
0.10	9.23E+01	-3.10E-05	-1.51E+02	4.54E+01	1.43E-02	-1.54E-03
0.15	9.81E+01	-3.45E-05	-1.65E+02	4.62E+01	1.72E-02	-1.84E-03
0.20	1.04E+02	-3.75E-05	-1.83E+02	4.70E+01	1.84E-02	-1.95E-03
0.25	1.07E+02	-3.74E-05	-1.87E+02	4.86E+01	1.83E-02	-1.94E-03
0.30	1.09E+02	-3.73E-05	-1.88E+02	4.99E+01	1.91E-02	-2.02E-03
0.35	1.12E+02	-3.86E-05	-1.97E+02	5.12E+01	2.00E-02	-2.12E-03
0.40	1.14E+02	-3.88E-05	-2.00E+02	5.21E+01	2.12E-02	-2.24E-03
0.45	1.16E+02	-3.95E-05	-2.06E+02	5.31E+01	2.24E-02	-2.37E-03
0.50	1.17E+02	-3.90E-05	-2.02E+02	5.48E+01	2.25E-02	-2.37E-03
0.55	1.19E+02	-3.87E-05	-1.98E+02	5.62E+01	2.34E-02	-2.47E-03
0.60	1.20E+02	-3.96E-05	-1.97E+02	5.76E+01	2.51E-02	-2.66E-03
0.65	1.25E+02	-4.22E-05	-2.09E+02	5.92E+01	2.62E-02	-2.79E-03
0.70	1.28E+02	-4.50E-05	-2.17E+02	6.09E+01	2.73E-02	-2.89E-03
0.75	1.33E+02	-4.73E-05	-2.25E+02	6.47E+01	2.73E-02	-2.93E-03
0.80	1.36E+02	-4.94E-05	-2.31E+02	6.79E+01	2.82E-02	-3.04E-03
0.85	1.44E+02	-5.36E-05	-2.49E+02	7.23E+01	2.87E-02	-3.13E-03
0.90	1.40E+02	-4.89E-05	-2.04E+02	7.82E+01	3.03E-02	-3.35E-03
0.95	1.51E+02	-5.72E-05	-2.09E+02	9.02E+01	3.00E-02	-3.42E-03
0.99	1.74E+02	-8.05E-05	-2.00E+02	1.12E+02	3.26E-02	-3.80E-03

Table 3: Parameters for the 3-parameter lognormal archetypal distributions of *BW* for each age/gender category for adults based on data from CSFII/NHANES III

Age category	Males				Females			
	$\square$	$\square_{LN(x)}$	$\square_{LN(x)}$	SS_{xy}	$\square$	$\square_{LN(x)}$	$\square_{LN(x)}$	SS_{xy}
20	34.815	3.618	0.349	1.28E-03	38.625	3.037	0.607	7.73E-04
21	34.313	3.659	0.339	1.09E-03	38.595	3.057	0.601	7.20E-04
22 – 24	33.229	3.732	0.321	8.43E-04	38.511	3.096	0.587	6.22E-04
25 – 29	31.038	3.849	0.294	6.45E-04	38.217	3.176	0.560	4.62E-04
30 – 34	28.159	3.961	0.270	6.48E-04	37.569	3.278	0.526	3.24E-04
35 – 39	24.997	4.051	0.250	7.30E-04	36.542	3.378	0.492	2.38E-04
40 – 44	21.560	4.127	0.234	8.02E-04	35.089	3.475	0.458	1.89E-04
45 – 49	17.351	4.201	0.217	8.30E-04	33.112	3.569	0.425	1.62E-04
50 – 54	12.629	4.269	0.202	8.00E-04	30.494	3.661	0.392	1.46E-04
55 – 59	0.011	4.426	0.171	7.41E-04	27.180	3.750	0.358	1.33E-04
60 – 61	0.000	4.418	0.171	6.39E-04	24.299	3.813	0.334	1.24E-04
62 – 64	0.000	4.410	0.171	5.79E-04	21.856	3.860	0.315	1.18E-04
65 – 66	0.000	4.401	0.170	5.39E-04	19.104	3.908	0.297	1.13E-04
67 – 69	0.000	4.390	0.170	5.33E-04	15.885	3.959	0.277	1.12E-04
70 – 74	-15.981	4.554	0.141	4.64E-04	9.217	4.056	0.243	1.23E-04
75 – 79	-26.074	4.630	0.126	6.60E-04	0.005	4.167	0.204	2.03E-04
80 – 84	-32.471	4.662	0.117	1.39E-03	-23.817	4.444	0.141	4.14E-04
85 – above	-46.227	4.766	0.102	2.15E-03	-30.735	4.495	0.125	8.11E-04